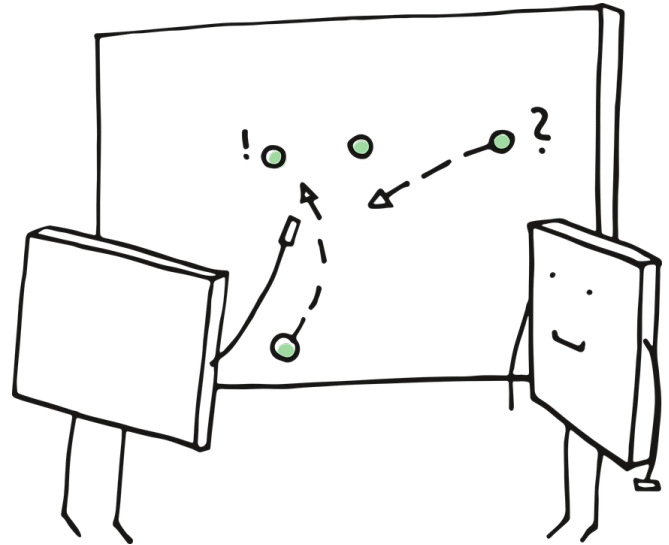
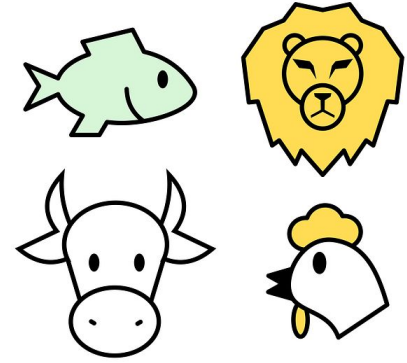


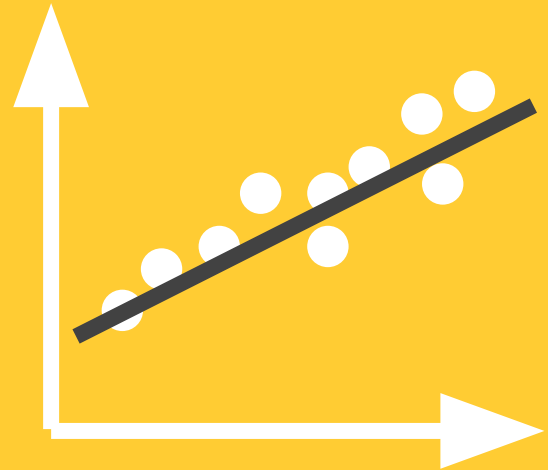
What is K Nearest Neighbours?



k-NN



- Used for classification and regression



k-NN

- Non-parametric, instanced-based learning algorithm

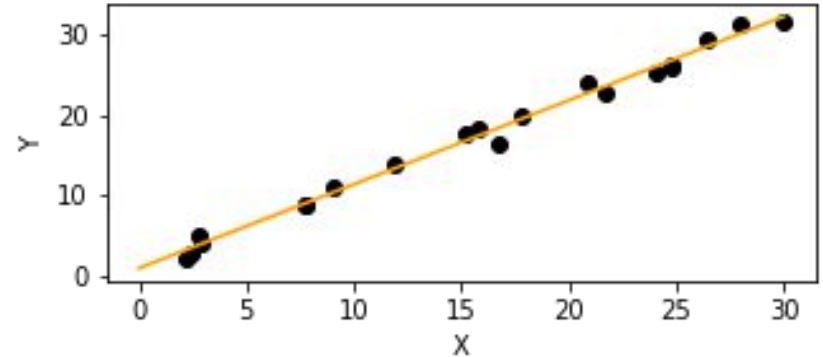
Non-Parametric

- Doesn't make explicit assumptions about the form of the mapping function
 - $h(X): X \rightarrow Y$

Parametric

- **Linear regression**

- $h(X): Y = wX + B$



Instance-Based

- Doesn't explicitly learn a model
- No training stage
- Training samples are only used during the prediction stage

k parameter

- Tuning parameter
- Controls the smoothness of the decision boundary of a k-NN classifier and its fit for a given dataset

Effects of k

Small k

- **k-NN**
 - high flexibility
 - sensitive to outliers
 - exhibits overfitting (low bias and high variance)
 - decision boundary is more jagged

Large k

- **k-NN**
 - More resilient to outliers
 - Lower flexibility
 - Exhibits underfitting (high bias and low variance)
 - Decision boundary is more smooth

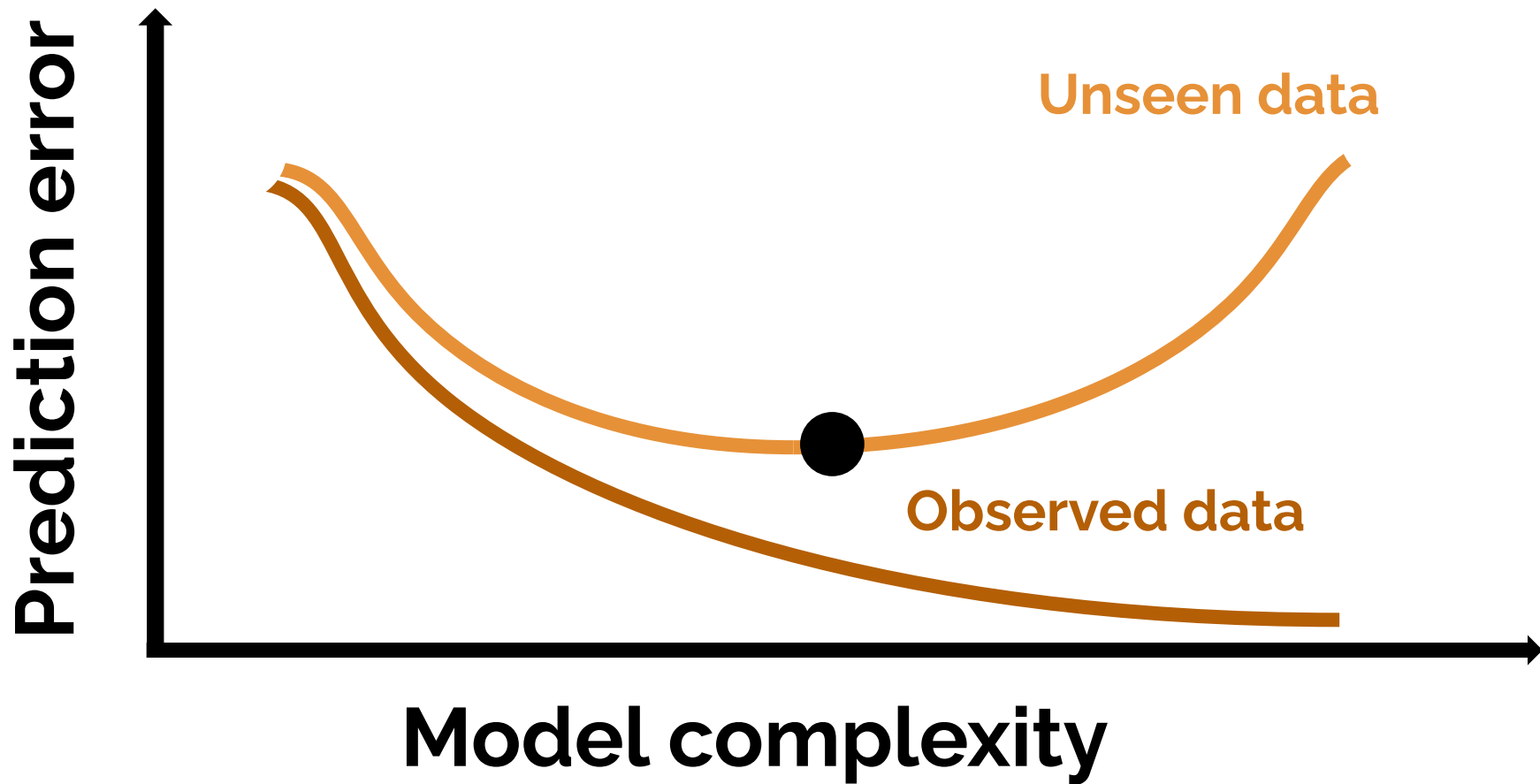
What is the best value of k ?

- Chosen using cross validation
- Select k that gives the lowest average cross-validation estimate of prediction error

**What is cross
validation?**

Central challenge in machine learning

- Algorithm must perform well on new inputs, not just training data



2 Goals of Machine Learning

- **Model selection**
- **Model assessment**

How do you choose a model? 🤔

- Find a model that fits future data best

How do you evaluate a model? 🤔

- Simplest approach: holdout cross-validation

Holdout cross-validation

- Randomly split data into:
 - Training set
 - Testing set

NEVER use test set to train

Holdout cross-validation 👎

- Fails to use all available data
- If $\frac{1}{2}$ of data is for training, poor hypothesis
- If $\frac{1}{10}$ of data is for testing, poor estimate of actual accuracy

What is the solution?

- **Cross validation**
 - Repeat training and validation on different randomly chosen subsets of original training set

K Fold cross validation

- **Each sample will be used for validation in 1 of the sets**
- **Each sample serves double duty: as training set and validation set**

How k-NN Works

k-NN Classification Step 1

- For each training sample, compute the distance between the test sample and the training sample

k-NN Classification Step 2

- Order the training samples by increasing distance

k-NN Classification Step 3

- **Select the first k training samples that have the smallest distance**
 - The k nearest neighbors of x

k-NN Classification Step 4

- Determine the predicted value $\hat{y}$ of x by the majority vote
 - (the most commonly occurring class of the k nearest neighbors)

k is usually an odd number to avoid ties

k-NN Pros

- Easy to implement
- Can achieve good classification performance when training set is large
- Works with multi-class datasets

k-NN Cons

- **Prediction stage is computationally intensive**
 - Algorithm has to iterate through all samples in the training set
 - Often an issue when the training set is large

k-NN Cons

- **High memory requirement**
 - Has to store all training samples all the time

k-NN Cons

- Unable to provide information as to whether a feature is more informative than another

What is K Nearest Neighbours? ✓

